Modeling/Computation

A26

Hierarchical Organization of Modularity in Metabolic Networks

Albert-László Barabási[1] (alb@nd.edu), Zoltán N. Oltvai[2] (zno008@nwu.edu), A. L. Somera[3], D. A. Mongru[3], G. Balazsi[3], Erzsebet Ravasz[1], S. Y. Gerdes[4], J. W. Campbell[4], and A. L. Osterman[4]

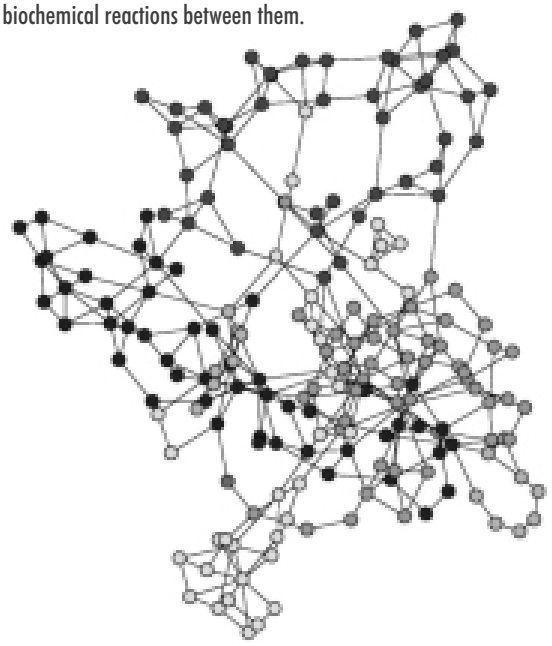
[1]University of Notre Dame, Department of Physics, 225 Nieuwland Science Hall, Notre Dame, IN 46556, 574-631-5767, Fax: 574-631-5952; [2]Department of Pathology, Northwestern University Medical School, Ward Bldg. 6-204, W127, 303 E. Chicago Ave., Chicago, IL 60611, 312-503-1175, Fax: 312-503-8240; [3]Northwestern University; and [4]Integrated Genomics, Inc.

The identification and characterization of system-level features of biological organization is a key issue of post-genomic biology. An elegant proposal addressing the cell's functional architecture is offered by the concept of modularity, assuming that the cell can be partitioned into a collection of modules. Each module, a discrete entity of several elementary components, performs an identifiable biological task, separable from the functions of other modules. Yet, it is now widely recognized that the thousands of components of the metabolism are dynamically connected to one another, such that the cell's functional properties are ultimately encoded into a complex metabolic web of molecular interactions. Within this network, however, modular organization and clear boundaries between sub-networks are not immediately apparent. Indeed, recent studies have demonstrated that metabolic networks have a scale-free topology. A distinguishing feature of such scale-free networks is the existence of a few hubs, highly connected metabolites such as pyruvate or CoA, which participate in a very large number of metabolic reactions. With a large number of links, these hubs integrate all substrates into a single, integrated web in which the existence of fully separated modules is prohibited.

To resolve the apparent contradiction, we now provided evidence that the metabolism has a hierarchical organization, an architecture that seamlessly integrates a scale-free topology with an inherent modular structure. For this purpose we have shown that the degree of clustering present in the network can be used as a distinguishing feature of a hierarchical structure, and offered direct evidence that the metabolism of 43 organisms have such a hierarchical architecture.

To turn this new conceptual framework into a practical tool we developed a method to directly identify and visualize the topological modules present in the *E. coli* metabolism and identified the function of these modules based on the predominant biochemical class of the substrates they belong to, using the standard, small molecule biochemistry based classification of metabolism. We find that most substrates of a given small molecule class are distributed within the same identified module and correspond to relatively well-delimited regions of the metabolic network, demonstrating strong correlations between shared biochemical classification of metabolites and the

Barabási— Fig. 1. The *E. coli* metabolic network color-coded based on the biochemical classification of the individual substrates. Each node corresponds to a metabolite, and links represent biochemical reactions between them.



global topological organization of *E. coli*. These results and the systematic experimental corroboration of this framework by global transposon mutagenesis will be discussed.

Supported by the DOE grant "The Organization of Complex Metabolic Networks." Principal Investigator: Albert-László Barabási, University of Notre Dame.

Background Literature

Hierarchical Organization of Modularity in Metabolic Networks, E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, *Science* Aug 30 2002: 1551-1555.

Experimental and System-Level Analysis of Essential and Dispensable Genes in *E. coli* MG1655, S.Y. Gerdes et al, in preparation.

A30

SimPheny: A Computational Infrastructure Bringing Genomes to Life

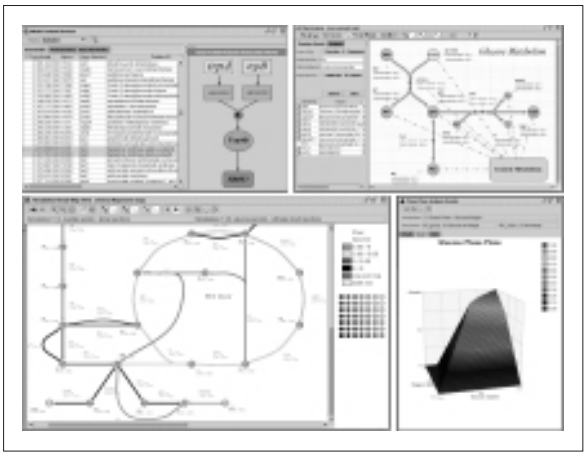
Christophe H. Schilling[1] (cschilling@genomatica.com), Radhakrishnan Mahadevan[1], Sung Park[1], Evelyn Travnik[1], Bernhard O. Palsson[2], Costas Maranas[3], Derek Lovley[4], and Daniel Bond[4]

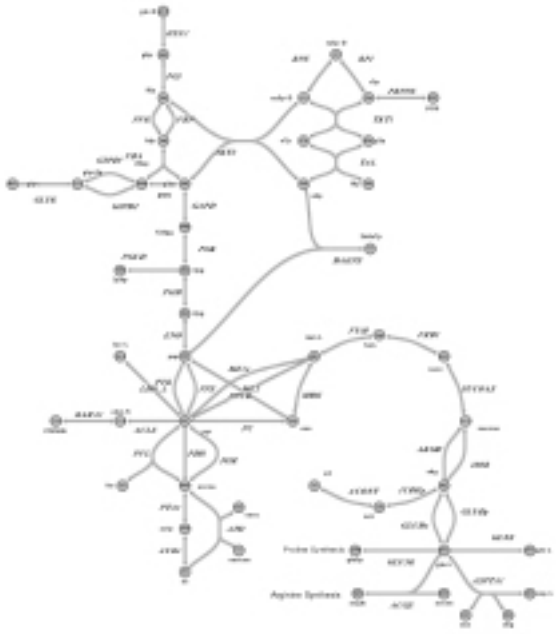
[1]Genomatica, Inc., 5405 Morehouse Drive , Suite 210, San Diego, CA 92121, 858-824-1771, Fax: 858-824-1772; [2]University of California, San Diego; [3]Penn State University; and [4]University of Massachusetts, Amherst

The Genomes to Life (GtL) program has clearly stated a number of overall goals that will only be achieved if we develop "a computational infrastructure for systems biology that enables the development of computational models for complex biological systems that can predict the behavior of these complex systems and their responses to the environment." At Genomatica we have developed the SimPheny™ (for Simulating Phenotypes) platform as the computational infrastructure to support a model-driven systems biology research paradigm. SimPheny enables the efficient development of genome-scale metabolic models of microbial organisms and their simulation using a constraints-based modeling approach.

We are currently utilizing this platform for a number of DOE-related projects including:

1. Developing the next generation of genome-scale models: In collaboration with Prof. Costas Maranas at Penn State University and Prof. Bernhard Palsson at the Univ. California, San Diego, we are integrating methods to incorporate regulation and signal transduction mechanisms into metabolic models and enable advance simulation algorithms that utilize mixed-integer linear programming (MILP).
2. *Geobacter sulfurreducens* Modeling: As part of the Microbial Cell Project led by Prof. Derek Lovley at the Univ. of Massachusetts we have completed the development of a first draft genome scale model for *G. sulfurreducens* within SimPheny. We are now beginning the process of performing simulations with the model to provide model-driven analysis of experimental data, and providing data integration solutions through the development of a model centric database
3. *Pseudomonas fluorescens* Model Development: As part of a Phase I Small Business Innovative Research (SBIR) grant we are constructing a genome scale model of *P. fluorescens* that will be used to drive metabolic engineering research on this organism for industrial bioprocessing applications.





Herein we will highlight the capabilities of the SimPheny platform as an infrastructure for supporting model driven systems biology research with special emphasis on its application to the development of the *G. sulfurreducens* metabolic model.

A32

Parallel Scaling in Amber Molecular Dynamics Simulations

Michael Crowley, Scott Brozell, and **David A. Case** (case@scripps.edu)

Dept. of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037

Large-scale biomolecular simulations form an increasingly important part of research in structural genomics, proteomics, and drug design. Popular modeling tools such as Amber and CHARMM are limited by both state-of-the-art hardware capabilities and by software algorithm limitations. Current macro-molecular systems of interest range in size to several hundred thousand atoms, and current simulations generally simulate one to tens of nanoseconds. With a 2 fs timestep, and each force evaluation involving millions of interactions to be calculated, a simulation requires many gigaflops to finish in a reasonable period of time. A parallel implementation of the calculation can provide the required performance by using the power of many processors simultaneously. However, communication speed between nodes has not progressed as rapidly as CPU processing power in recent years. Here, we address some weakness of the current parallel molecular dynamics implementation in Amber (and in a comparable program such as CHARMM). The work is aimed at making affordable a new generation of increasingly sophisticated biomolecular simulations.

Atom-Based Decomposition in Amber

Of the many ways to distribute the work of a force calculation in parallel [1,2], the method of replicated data (or "atom decomposition") has traditionally been used in Amber and CHARMM. This sort of parallel implementation is based on dividing each portion of the force calculation evenly among the processors, while keeping a full set of coordinates on all processors. This is very flexible, and relatively straightforward to program. Each processor is assigned an equal number of bonds, angles, dihedrals, and nonbond interactions. In this way, the work is balanced in each part of the force calculation, and the computation time scales well as the number of processors increases. However, in each part of the force calculation a node computes forces for different subsets of atoms. For this reason, each processor requires a complete set of up-to-date coordinates and is assumed to have components of forces for all atoms. At each step, the forces computed for all atoms on each node must be summed and distributed, and updated coordinates must be collected from each node and sent complete to all nodes. There are hence two all-to-all communications at each step. Even with binary tree algorithms for distributed sums and redistribution, the communication time becomes a significant fraction of the total time by 32 processors, even on the most sophisticated parallel machines. This limitation eliminates the possibility of efficient parallel runs at large numbers of processors, and puts a restriction on the size and length of simulations that a researcher can attempt even when large parallel computational resources are available. Still, for systems up to about 32 processors, these codes are more efficient for typical solvated simulations than are popular alternatives such as CHARMM or NAMD.

Spatial Decomposition in Amber

The second-generation parallel Amber, now under development, implements a "spatial decomposi-

tion" method [1,2] in which the molecular system is divided into regions of space where approximately equal amount of force computation is required. The method works when contributions to the force on an atom come primarily from interactions with other atoms that are relatively close and are neglected for atoms that are beyond a fixed cutoff. (This condition is valid in modern MD simulations except for long-range electrostatics, which use Ewald-based methods discussed below.) In this approach, a processor is assigned the atoms located in a slice of space and it is responsible for the coordinates, forces, velocities, and energetic contributions of those atoms. In order to compute the forces for its *owned* atoms, the processor must be able to compute the contributions from interactions with atoms that are within the cutoff, including any that are assigned to other processors. A processor keeps a copy of all such *needed* atom coordinates and forces as well as its *owned* atom coordinates and forces. At each step, a processor determines the force contributions due to all interactions in its owned and needed atoms. It sends all force contributions on *needed* atoms to the processors that own those atoms and receives any force contributions for its *owned* atoms that were calculated by other processors. When the force communications are complete, the coordinate integration is performed on the *owned* atoms. Each message in all the above communications is at most the size of the *owned* atom partition and will often be considerably smaller. Inventories of message sizes shows a reduction in overall data transferred to less than half of that for the replicated-data method, fort typical solvated protein or nucleic acid systems. Timing for communication is reduced by nearly identical ratios.

This conversion of the Amber codes is complex, since there are complications inherent in spatial decomposition that do not arise in the replicated data method; these are mainly in the treatment of bonded interactions, constraints, long-range electrostatics, and bookkeeping. The first two complications arise when molecules (chemical bonds) or distance constraints span the spatial boundaries. Most bonds, angles, dihedrals, restraints, and constraints can be assigned according to ownership of atoms. When the atoms involved are owned by distinct processors, an algorithm must be implemented to insure that the interactions are considered but only once, and that the coordinates necessary are current and correct. Bond-length constraints (using the so-called "SHAKE" approach) are more complicated, since they redefine the positions of atoms after the computed forces have been applied to owned atoms. In this case, the updated positions of all atoms involved in a constraint must be known in order to adjust positions of owned atoms regardless of whether they are owned or not. Besides these complications lie the bookkeeping needed to keep track of which forces and coordinates are being sent and received. One of the challenges is to keep the bookkeeping to a minimum, and to make it as efficient as possible, so that it does not simply replace the time saved in communications. Finally, we must optimize scaling of the Ewald method of treating long-range electrostatics in periodic system, and in particular, the PME implementation of Ewald sums. We are currently exploring several methods of reducing the communications costs of PME in highly parallel systems.

Running Dynamics on Pentium Clusters

In molecular dynamics simulations the calculations of the nonbonded interactions are a computational bottleneck. These interactions depend upon the interparticle distances. On Intel 32 bit architectures (IA32) the fastest methods to evaluate reciprocal square roots utilize the Streaming SIMD (Single-instruction multiple-data) Extensions for double precision (SSE2) operations. Tuning the AMBER source code to enable automatic vectorization, i.e., generation of SSE2 instructions by compilers, introduces a memory cache penalty as well as additional loop overhead. On IA32 platforms the SSE2 tuned AMBER is approximately eight percent faster than the original AMBER as measured by execution times of a typical explicit solvent protein simulation. The IA32 SIMD vector length is four for single precision data and two for double precision. Conversion of AMBER to single precision yields a forty percent performance improvement on IA32 platforms, as measured by execution times of a typical Generalized Born implicit solvent protein simulation. We are exploring what how to make best use of these sorts of gains without sacrificing any essential precision in the results.

[1] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117,** 1-19 (1995).

[2] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **151**, 283-312 (1999).

A34

Microbial Cell Model of *G. sulfurreducens*: Integration of in Silico Models and Functional Genomic Studies

Derek Lovley[1] (dlovley@microbio.umass.edu), Maddalena Coppi[1], Daniel Bond[1], Jessica Butler[1], Susan Childers[1], Teena Metha[1], Ching Leang[1], Barbara Methé[2], Carol Giometti[3], R. Mahadevan[4], C. H. Schilling[4], and B. Palsson[4]

[1]Department of Microbiology, University of Massachusetts, Amherst, Amherst, MA; [2]The Institute for Genomic Research, Rockville, MD; [3]Biosciences Division, Argonne National Laboratory, Argonne, IL; and [4]Genomatica, Inc., San Diego, CA 92121

Molecular analyses have demonstrated that *Geobacter* species are the predominant dissimilatory metal-reducing microorganisms in a variety of subsurface environments in which metal reduction is an important process, including uranium-contaminated aquifers undergoing bioremediation. The long-term objective of this study is to develop comprehensive conceptual and mathematical models of *Geobacter* physiology and its interactions with its physical-chemical environment, in order to predict the behavior of *Geobacter* in a diversity of subsurface environments. Prior to sequencing and initiating the functional analysis of the genomes of *Geobacter sulfurreducens* and *Geobacter metallireducens*, it was considered that these organisms were non-motile, strict anaerobes with a simple metabolism that required little regulation. It is now clear that each of these basic characterizations is incorrect. These and other surprises from the analysis of the *Geobacter* genomes are significantly influencing our design of strategies for *in situ* metals bioremediation.

A preliminary genome-scale metabolic model of *G. sulfurreducens* central metabolism has been developed with a constraints-based modeling approach using the annotated genome sequence and scientific literature. A detailed description of this model is provided in the companion poster presented by Genomatica.

Further experimental investigation is required to refine and expand this preliminary *in silico* model. For example, *G. sulfurreducens* requires a fumarate reductase in order to grow with fumarate as the sole electron acceptor, but when growing with Fe(III) as an electron acceptor, it also requires a succinate dehydrogenase in order to complete the oxidation of acetate to carbon dioxide via the TCA cycle. Although the genome of *G. sulfurreducens* contains a cluster of three genes, *frdA, frdB*, and *frdC*, resembling the three subunits of the *Wolinella succinogenes* fumarate reductase, genes for a separate succinate dehydrogenase are not apparent. Studies with a *frdA* knock-out mutant revealed that this complex functioned *in vivo*, not only as a fumarate reductase, but also as a succinate dehydrogenase. Elucidation of this novel function significantly improves the *in silico* model.

A genetic study on putative hydrogenase genes revealed that although *G. sulfurreducens* has at least four possible hydrogenases, only one is required for growth with hydrogen as the sole electron acceptor. Interestingly, this is the one hydrogenase gene that is missing from the closely related *Geobacter metallireducens*, which unlike *G. sulfurreducens* can not grow with hydrogen as an electron donor. Genetic and metabolomic studies have demonstrated that a novel cytoplasmic enzyme, which had previously been identified as NADPH-dependent Fe(III) reductase, is in fact involved in electron transport from NADPH to a variety of electron acceptors and plays a key role in intracellular NADPH homeostasis. In addition to providing valuable information on *Geobacter* metabolism, this result underscores the importance of investigating the function of enzymes *in vivo*, especially when dealing with electron transport to Fe(III), which many enzymes can nonspecifically reduce *in vitro*.

The *Geobacter* genomes contain a much higher percentage of genes devoted to electron transport function than found in other organisms. Gene expression and proteomics studies have revealed specific genes that are associated with Fe(III) reduction. For example, genes with high homology to flagellar and type(IV) pilus genes were specifically expressed during growth on the insoluble Fe(III) and Mn(IV) oxides. This finding, coupled with the discovery that *Geobacter* is chemotactic to Fe(II), has suggested that *Geobacter* produce extracellular appendages for motility only when they need to search for insoluble Fe(III) or Mn(IV) oxides (Childers, S. E., S. Ciufo, and D. R. Lovley. 2002. *Geobacter metallireducens* access Fe(III) oxide by chemotaxis. *Nature* 416:767– 769). Studies on mutants missing key pilin genes have further demonstrated the important role of pili in Fe(III) oxide reduction. In a similar manner, genes that are homologs of

components of the type II secretion systems of other bacteria are specifically expressed during growth on Fe(III) or Mn(IV) oxide and disrupting these genes inhibits the reduction of the insoluble metal oxides, but not soluble electron acceptors, including Fe(III) citrate. Mutants with deletions in one of several *c*-type cytochrome genes that are more highly expressed during growth on Fe(III) no longer had the capacity for Fe(III) reduction. These studies clearly demonstrate that *Geobacter* possess a highly regulated system of genes that are specifically involved in the reduction of Fe(III) and Mn(IV).

Additional examples, of ongoing iterative modeling and experimental elucidation of metabolic and electron transport pathways will be presented. These include novel enzymes for carbon metabolism as well as the surprising finding, originating from genome-based modeling, that *Geobacter* species that were previously classified as "strict anaerobes" have the ability to use oxygen. This has important implications for their survival under the fluctuating redox regimes common in the subsurface. Genome analysis has also suggested additional bioremediation capabilities such as the ability to degrade TNT and related contaminants and the ability to reduce mercury. These early results under the Microbial Cell Project demonstrate the power of genomic analysis to significantly influence bioremediation research.

A36

Towards a Self-Organizing and Self-Correcting Prokaryotic Taxonomy

George M. Garrity[1] (garrity@msu.edu) and Timothy G. Lilburn[2]

[1]The Bergey's Manual Trust, Michigan State University, East Lansing, MI 48224; and [2]The American Type Culture Collection, Manassas, VA 20110

A longstanding goal of microbiologists has been the creation of a taxonomy that reflects the natural history of prokaryotes and provides a stable and reliable scheme that is predictive and workable in a wide variety of applications. For genomic comparisons the establishment of such a framework is essential if we hope to be able to recognize homologous, paralogous and xenologous sequences. Over the past 15 yrs, a reasonably good picture of the evolutionary relationships among the *Bacteria* and *Archaea* has emerged, largely as a result of the widespread adoption of 16S rRNA as the molecule of choice for phylogenetic studies. At present, two large-scale phylogenetic trees of the prokaryotes exist in varying states of completeness and serve as the foundation of the comprehensive taxonomy used in the *Second Edition of Bergey's Manual of Systematic Bacteriology*. However, limitations of the underlying phylogenetic models preclude incorporation of much of the available sequence data into an all-inclusive taxonomy that can be easily maintained and modified in an automatic fashion as new taxa are described and existing taxa emended. Confounding problems include a high number of annotation errors in the sequence data and a lack of clear criteria for defining taxon boundaries.

We recently described the application of techniques drawn from the field of exploratory data analysis that take advantage of the large number of SSU rRNA sequences available and that produce results which are reconcilable with our knowledge of both the phylogenetic models and available phenotypic data. We found that principal components analysis (PCA) of large matrices of evolutionary distance data ($> 2\times10^6$ data points) yielded 2D and 3D maps of evolutionary space in which the high level groups that were formed proved consistent with those found in the large scale consensus trees. While PCA maps proved useful in establishing a comprehensive prokaryotic taxonomy and greatly aided in the identification of misclassified sequences, the method proved less useful in establishing group membership of misclassified or unknown sequences. In order to eliminate some of these problems, we have turned to a second visualization technique, heat maps. Heat maps are a type of graph in which signal intensity is displayed as a gradation of color within the confines of a precisely ordered grid. Heat maps have recently found widespread application in the field of microarray analysis and provide a useful means of visualizing differential gene expression. Heat maps, in an earlier variation, have also found application in the distant past in prokaryotic taxonomy.

Recently, we have begun using heat maps as a graphical tool for comparing alternative phylogenies and models at intermediate taxonomic levels (Order-Genus). In this paper, we describe how heat maps were used as a graphical tool to demonstrate significant improvements in the current classification of the *Gammaproteobacteria,* brought about by the application of a newly developed supervised-clustering

algorithm. Using a set of simple statistical criteria for group membership, the algorithm iteratively reorganizes the distance matrix on a taxon-by-taxon basis, excludes outliers and misidentified sequences, and subsequently reinserts such sequences into the matrix at the location of its most-probable identity. Dynamically reordered heat maps of user-selected submatrices serve as an aid to the inspection and modification of the automatically generated classification, the identification of possible classification errors, ad-hoc testing of alternative classifications/hypotheses and direct extraction of the underlying distance data.

Our results demonstrate that significant improvements to prokaryotic taxonomy can be readily obtained using statistical approaches to the evaluation of sequence-based evolutionary distances. Errors in curation, classification, and identification can be easily spotted and their effects corrected, and the classification itself can be modified so that the information content of the taxonomy is enhanced. Furthermore, evolutionary analyses based on other molecules can be viewed in terms of this rRNA-based phylogeny and used to improve the classification in taxa where the information content and resolving power of the 16S rRNA molecules proves too low (e.g., *Bacillus*). Obviously, a more robust classification has greater predictive power and serves as a more reliable evolutionary framework for genome exploration. The visualization supplies an intuitive approach, in that persons with no taxonomic training can, by looking at the heat maps, see how the classification might be improved; our algorithm formalizes and automates the means used to achieve such improvements. The chief drawback of the approach is that groups formed from taxa that are sparsely represented in the SSU rRNA sequence data set may not be as robust or stable as groups from more richly populated taxa. This is especially true in instances where such taxa are equidistant to two or more otherwise unrelated taxa. However, such problems should prove transitory as the data set grows daily and the tools we are developing will allow us to maintain and expand a comprehensive prokaryotic taxonomy.

It is worthwhile noting that the usefulness of the techniques described here is not limited to bacterial taxonomy. These methods can be used to develop and improve classifications of all types. For example, functional assignment of new sequences benefits from a reliable protein classification. Data from gene expression microarrays might also be usefully classified using these techniques.

A pseudocode description of the methods used here will be available at the poster.

A38

Computational Framework for Microbial Cell Simulations

Haluk Resat[1] (haluk.resat@pnl.gov), Heidi Sofia[1], Harold Trease[1], Joseph Oliveira[1], Samuel Kaplan[2], and Christopher Mackenzie[2]

[1]Pacific Northwest National Laboratory and [2]University of Texas Medical School at Houston

The complex nature of data characterizing cellular processes makes mathematical and computational methods essential for interpreting experimental results and in designing new experiments. Although the need to develop comprehensive approaches is widely recognized, significant improvements are still necessary to bridge experimental and computational approaches. Such advances require the development of integrated sets of computational tools to achieve the level of sophistication necessary for understanding the complex processes that occur in cells. As part of this project, we have been developing a set of prototype network analysis tools and methods, and employing them to investigate the flux and regulation of fundamental energy and material pathways in *Rhodobacter sphaeroides*.

Our tool development efforts span a wide spectrum. Current prototype components are designed in such a way that, when combined later, they will form the backbone of a comprehensive microbial cell simulation environment. In particular, we have been working on developing a framework for the following areas:

Qualitative network analysis: This analysis algorithm uses the connection diagram of the cellular networks to classify and rank them according to their linkage characteristics. We have shown that the Petri net representation can be a powerful way to extract the topologies and the universal features of cellular networks. This allows for the comparison of networks to decipher the common modules.

Gene regulatory networks: We have developed an object oriented stochastic simulation software that can be used to study the expression levels of genes. Given a regulatory network and using a user defined multistate representation for gene

expression levels, our software can be used to simulate the mean gene expression levels and their fluctuations. This simulation software was applied to derive the network parameters of a recently reported synthetic genetic network.

Stochastic kinetic simulation software: It has been well established that the number of molecules of certain species in cells can be very low. This makes the stochastic representation more appropriate to study the dynamics of cellular systems. We have shown that the kinetic simulations are not only limited to biochemical reactions but any physical event such as vesicle formation can be included in kinetic simulations. Our kinetic simulation algorithm was devised and implemented in such a way that it can be used to simulate hybrid models that combine biochemical and physical events.

Imaging of bacterial cells and image recontruction: We are in the process of obtaining images of *R. sphaeroides* using electron tomography. *R. sphaeroides* cells will be chemically processed for TEM and imaged using the high resolution electron microscope at EMSL/PNNL. A different method of electron tomography will be utilized for 3-dimensional reconstruction of a bacterial cell to visualize the spatial distribution of intracellular vesicles throughout the bacterium. This will be performed using the remove operation capabilities of the TEM imaging facility at UCSD that allows us to obtain a series of tilted digital images that can later be computationally reconstructed. Reconstructed 3-D geometry of surface features and of the internal structures will later be used in spatially resolved simulation studies.

Mesh grid based simulation framework: Explicit incorporation of geometric and structural information into cellular models is important to study the effect of the local environment on transport and kinetic properties. NWGrid and NWPhys codes that are part of the large scale computational simulation framework used at PNNL have been adapted to biological systems. Obtained images of *R. sphaeroides* will define a structure upon which we can map spatially dependent quantities and will provide the basis for building spatial computational models of this microbial cell.

Similarity analysis of genomes and prediction of superfamilies: We have developed analysis and visualization software based on clustering and data integration to enable biologists to navigate through large quantities of genome sequence data and operon information for the purpose of classifying genomic sequences and assigning protein functions rapidly and efficiently. We are applying the Similarity Box analysis to *R. sphaeroides* genome sequence data. To illustrate the usefulness of our software, we show a comparative genomics analysis of the FNR superfamily, an important group of transcriptional regulator proteins in diverse bacterial species that respond to signals such as redox, nitrogen status, and temperature. We also show a genomic comparison of FNR proteins in the two ?-Proteobacterial species, *R. sphaeroides* and *Rhodopseudomonas palustris.* These two closely related but divergent prokaryotes have a partially overlapping complement of FNR proteins.

Molecular part list and network derivation: Although most of the key genes have been identified, the network describing the energy metabolism of *R. sphaeroides* is minimally understood. To improve the energy metabolism network of *R. sphaeroides* that was initially built using biochemical data, we are using the recent genome (DOE sponsored JGI) and microarray (UT-Houston) data to build a more complete molecular parts list. We are making use of the genome data by applying our similarity analysis approach to assign functionality to improve the annotation of the genome. For similar purposes, we are analyzing the microarray data using clustering methods combined with promoter sequence information.

A40

Characterization of Genetic Regulatory Circuitry Controlling Adaptive Metabolic Pathways

Harley McAdams*, Lucy Shapiro*, and Mike Laub*

*Presenters

Stanford University

In this project, an interdisciplinary team of scientists from Stanford, Harvard, and SRI International is characterizing genetic regulatory circuitry and metabolic pathways in the bacterium *Caulobacter crescentus*. The *Caulobacter* are oligotrophic organisms adapted to low-nutrient environments such as clear streams, lakes, and the open ocean and some species are found in the deep subsurface environment. The genetic circuit controlling the *Caulobacter crescentus* cell cycle is

one of the best-characterized bacterial regulatory networks. *Caulobacter* is a particularly useful model system for study of cell-cycle regulation because cell populations can be synchronized with minimum perturbation of the normal physiology of the cell. This feature has permitted a detailed molecular analysis of the *Caulobacter* life cycle and has revealed a complex regulatory network governing the cell cycle progression and morphogenesis. The results of the current project will provide a powerful base for eventually engineering situation-specific regulatory "cassettes" into *Caulobacter* cells for targeted remediation applications.

In the first year of the project, we have developed a computational method to predict *Caulobacter* operon organization, established a baseline profile of gene expression levels for the complete genome over the cell cycle using microarrays, and developed and optimized a protocol for high throughput creation of gene deletion mutants for the entire *Caulobacter* genome. We have succeeded in obtaining and visualizing unique biofilms of *Caulobacter* wild-type and mutant cells. In their natural habitat, *Caulobacter* commonly grow in biofilms. We are particularly interested in determining whether genes are expressed in biofilms that are not active under laboratory conditions. As with most newly sequenced genomes, about forty percent of *Caulobacter*'s genes are of unknown function. Determining what these genes do is an important objective of the project.

We have published the first version of the *CauloCyc* online database (www.biocyc.org) for browsing and analysis of the *Caulobacter* genome using SRI's Pathway Tools software. This combined database and software environment has powerful and unique capabilities for modeling, visualizing, and analyzing biochemical and genetic networks. Using SRI's PathoLogic program we have computationally predicted *Caulobacter's* metabolic pathways from the genome. The PathoLogic program accepts two inputs: a fully annotated genome sequence, and the MetaCyc metabolic pathway database. The process of pathway prediction involves evaluating the evidence for the presence of each pathway in the reference DB for the organism being analyzed. A pathway consists of a sequence of enzyme-catalyzed reactions. The more enzymes we find within a pathway that are encoded by the genome, the more evidence we say there is for the presence of that pathway. The resulting metabolic pathway prediction assigned 617 *Caulobacter* enzymes to their corresponding metabolic reactions in 130 predicted metabolic pathways. Now we are investigating the "holes" within these predicted metabolic pathways, and we are using microarray studies of colonies growing in diverse media plus focused Blast studies to identify the missing enzymes within the *Caulobacter* genome.

A28

Computational Elucidation of Metabolic Pathways

Imran Shah (imran.shah@uchsc.edu)

School of Medicine, University of Colorado
http://shah-lab.uchsc.edu

Elucidating the metabolic network of a living system is an important requirement for modeling its physiological behaviour and for engineering its pathways. With the availability of whole genomes it is theoretically possible to infer the presence of putative enzymes and transporters in an organism. However, piecing this information into a complete picture is still mostly a daunting manual task for at least two reasons. First, we do not have accurate and sufficient annotation of enzymatic function from sequence. Consequently, many proteins in completely sequenced microbes remain functionally uncharacterized. Second, inferring the causal biochemical connections within a metabolic network is not straightforward. We are developing a computational infrastructure to address these challenges. In earlier work we have developed a machine learning (ML) approach to improve the assignment of enzymatic function from sequence. More recently, we have developed an artificial intelligence (AI) approach for the prediction of metabolic pathways and their interactive visualization, called PathMiner. This poster will present an overview of our system and discuss some relevant results for the radiation resistant microbe, *D. radiodurans*, and the metal-reducing bacterium, *S. oneidensis* MR-1.

A42

Data Exchange and Programmatic Resource Sharing: The Systems Biology Workbench, BioSPICE and the Systems Biology Markup Language (SBML)

Herbert M Sauro (hsauro@kgi.edu)

Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711

Standards

There is now a wide variety of modeling tools available to the budding systems biologist, but until recently there has been no agreed way to exchange models between different tools. The Systems Biology Markup Language is one of two emerging standards to allow systems biology modeling and analysis packages to exchange models. SBML is an open XML-based format developed to facilitate the exchange of models of biochemical reaction networks between software packages. Currently SBML Level 1 is supported by a growing number of tools, including, Jarnac, JDesigner, Gepasi, VCell, jigCell, Cellerator, and BioSPICE (under development). Level 2 is currently under discussion with the BioSPICE group, with a final release sometime in the second quarter of 2003.

There are in addition a small number of growing repositories of SBML models now available on the web, including, http://www.sbw-sbml.org, http://www.symbio.jst.go.jp/~funa/kegg/mge.html, http://www.sys-bio.org, http://www-aig.jpl.nasa.gov/public/mls/cellerator/nb.html, http://www.gepasi.org/

Programmatic Resource Sharing

The ERATO Systems Biology Workbench (SBW) is an open source, portable (Windows, Linux, Mac OS X) framework for allowing both legacy and new application resources to share data and algorithmic capabilities. Our target audience is the computational biology community whose interest lies in simulation and numerical analysis of biological systems. SBW allows communication between processes potentially located across a network on different hardware and operating systems. SBW currently has bindings to C, C++, Java, Delphi, Python, Perl and BioSPICE, with more planned for in the future. SBW uses a simple messaging system across sockets as a means for applications to communicate at high speed.

Software components that implement different functions (such as GUIs, model simulation methods, analysis methods, database interfaces, etc.) can be connected to each other through SBW using a straightforward application programming interface (API). The figure illustrates the visual design tool, JDesigner, interacting with the computational engine, Jarnac, via SBW (Jarnac is acting as a server and is thus invisible to the user). This setup permitted us to avoid writing, 'yet-another-simulator', and allowed JDesigner, which had no inherent simulation capabilities to dispatch simulation requests to another resource. Under SBW, models are exchanged between resource applications using SBML.

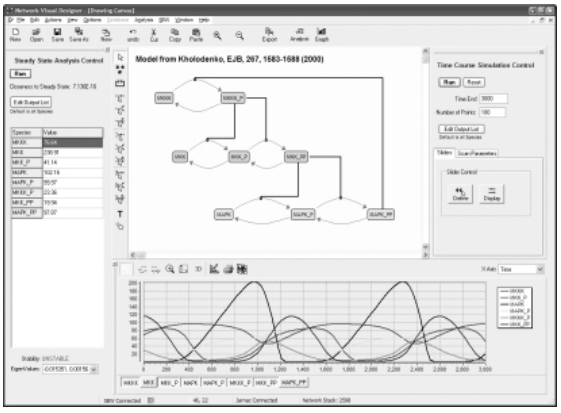
We are also working closely with the BioSPICE DARPA funded program to enhance SBML to Level 2 and to establish a set of agreed resource APIs to enable plug and play resources for both BioSPICE and SBW.

There is a growing list of modules which can communicate with each other via SBW and the BioSPICE software, including time course simulators, stochastic simulators (at least three kinds), basic optimizer, structural analysis tool via METATOOL, graphing tools, system browsing tools, etc.

Collaborators

The development of SBML/SBW was primarily conducted at Caltech by Hiroaki Kitano, John Doyle, Hamid Bolouri, Mike Hucka and Andrew Finney and Herbert Sauro in collaboration with

Sauro - Fig. 1. JDesigner simulating a MAPK Pathway



many other groups, including (in alphabetical order): A. Arkin, B. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, D. Fell, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. Juty, J. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. Loew, D. Lucio, P. Mendes, E. Mjolsness, Y. Nakayama, M. Nelson, P. Nielsen, T. Sakurada, J. Schaff, B. Shapiro, T. Shimizu, H. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, the BioSPICE program project investigators.

Web sites for further information on SBW and SBML:

- www.sbw-sbml.org
- www.sys-bio.org

Funding was provided by ERATO, the Keck Graduate Institute, DARPA, and a U.S. Air Force Grant.

A44

A Web-Based Laboratory Information Management System (LIMS) for Laboratory Microplate Data Generated by High-Throughput Genomic Applications

James R. Cole[1] (colej@msu.edu), **Joel A. Klappenbach**[1] (klappenb@msu.edu), Paul R. Saxman[1], Qiong Wang[1], Siddique A. Kulam[1], Alison E. Murray[2], Liyou Wu[3], Jizhong Zhou[3], and James M. Tiedje[1]

[1]Center for Microbial Ecology, Michigan State University, East Lansing, MI; [2]Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV; and [3]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

The use of high-throughput genomic-level methodologies such as microarray fabrication, proteomics, and whole-genome clone libraries necessitate computer database management tools for tracking and archiving massive quantities of data. We have developed a laboratory information management system (LIMS)—named the MicrobeArrayDB—for handing high-throughput data created during the fabrication of DNA microarrays. The laboratory microplate (96 or 384 well) functions as the primary data unit of the LIMS. This flexible database structure permits creation of different plate types, with associated data fields, extending the functionality of the LIMS to many different applications. Project-specific customization of the LIMS is controlled through a set of easily modified meta-data tables containing information on microplate types, contents, and how contents of microplates are combined and stored during laboratory procedures. The contents of microplates ("reagents") serve as "reactants" that are combined by the user during *in silico* "reactions" to create new product plates within the database. Users load microplate-specific information using cut-and-paste operations from tab-delimited file formats such as those created during oligonucleotide primer/probe design and from manufacturer supplied files. Data from these files is inherited during subsequent "reactions" that create new microplates. LIMS tools are interfaced through any internet browser and data access is controlled through group and user level permissions. User name and time stamps are recorded for each entry into the LIMS creating a permanent record and detailed audit trail. The MicrobeArrayDB is built on a multi-tier client/server architecture model using a publicly-available relational DBMS for our back-end tier (PostgreSQL) and Java web technologies for middle and presentation tiers.

In our initial project configuration, the LIMS is customized to track the production of a PCR-based DNA microarray from primer design to product deposition on coated slides. Current plate types for the microarray configuration include: Primer, Template, Primer Pair, PCR Product, and Printing plates. Contents of these plates are combined during "reactions"—such as the creation of a PCR Product plate from existing Primer and Template plates—as they are physically combined in the laboratory. Data inheritance is structured with microplate data loaded from text files produced during primer design and synthesis that include information such as gene ID/name, primer sequences, design criteria, quantity, length, batch control numbers, etc. Process information such as PCR reactions, product purification, and quality control data are added by the user during microarray construction. Internal testing has been conducted to track the fabrication process of whole-genome expression arrays for *Shewanella oneidensis* MR-1 at Oak Ridge National Laboratory. This implementation of the LIMS is targeted for whole-genome microarrays for bacteria including *Deinococcus radiodurans R1, Desulfovibrio vulgaris*, *Geobacter metallireducens*, *Nitrosomonas europaea, Rhodopseudomonas palustris.*

A46

BioSketchpad: An Interactive Tool for Modeling Biomolecular and Cellular Networks

Jonathan Webb[1], Lois Welber[1], Arch Owen[1], **Jonathan Delatizky**[1] (delatizky@bbn.com), Calin Belta[2], Mark Goulian[2], Franjo Ivancic[2], Vijay Kumar[2], Harvey Rubin[2], Jonathan Schug[2], and Oleg Sokolsky[2]

[1]BBN Technologies (http://bio.bbn.com/) and [2]University of Pennsylvania

The Bio SketchPad (BSP) is an interactive tool for modeling and designing biomolecular and cellular networks. It features a simple, easy to use, graphical front end. Descriptive models can be built and parameterized, and converted to forms supported by external simulation and analysis tools. The current version of BSP supports CHARON, a high level language and toolset for simulating hybrid systems developed at the University of Pennsylvania, and is being enhanced to communicate with other simulators, such as Virginia Tech's JigCell.

BSP was designed with biologists for biologists. It provides an intuitive graphical interface which allows experimentalists to easily generate working models of networks. Biomolecular reactions supported include transcription, translation, regulation, and general protein-protein interactions. BSP supports typical editing operations for graph editors as well as some specialized operations specific to the biomolecular modeling domain. Chemical species, reactions, and regulations of reactions are drawn as nodes in the graph. The user can control some of the rendering properties of specie elements such as the text label, color, and shape of the drawn node. Syntactic constraints are imposed on the model construction. Node highlighting is used to assist users during model construction. Specialized commands are implemented for constructing reaction geometries common to biochemical systems with specified numbers of inputs and outputs. Model parameters are specified in parameter dialogs accessed through the graphical presentation of the model. The set of parameters can change depending on which types of rate laws or regulation functions are used.

BSP is under active development. Recent and upcoming enhancements include the use of SBML Level II to exchange model information with other in silico tools, modularization of the simulator interface to utilize the standards being developed in the DARPA IPTO BioComp program, and the ability to represent models hierarchically.

BSP development has been funded by the DARPA IPTO BioComputation Program, PM Dr. Sri Kumar. The BSP application together with the CHARON simulation environment are available for download.

A48

Molecular Docking with Adaptive Mesh Solutions to the Poisson-Boltzmann Equation

Julie C. Mitchell[1] (mitchell@sdsc.edu), Lynn F. Ten Eyck[1], J. Ben Rosen[2], Michael J. Holst[3], Victoria A. Roberts[5], J. Andrew McCammon[4], Susan D. Lindsey[1], and Roummel Marcia[1]

[1]San Diego Supercomputer Center, [2]Department of Computer Science and Engineering, [3]Department of Mathematics, [4]Department of Chemistry and Biochemistry and Department of Pharmacology, University of California San Diego; and [5]Department of Molecular Biology, The Scripps Research Institute

The Docking Mesh Evaluator (DoME) uses adaptive mesh solutions to the Poisson-Boltzmann equation to quickly evaluate and optimize docking energies. This is accomplished by interpolation of potential functions over an irregular mesh that is dense in high gradient regions. The result is a method capable of performing detailed energy calculations very quickly. The initial version of DoME offers many useful tools for computational study of molecular interactions. DoME is intended to bridge the gap between methods that use a coarse interaction model for computational efficiency and those having detailed but expensive calculations. The software is fully parallel and can run on supercomputers, clusters and linked independent workstations.

The Critical Assessment of PRedicted Interactions (CAPRI) is a CASP-inspired exercise in which the goal is to predict bound protein-protein structures given their individual crystal structures. Using the Fourier transform-based molecular program DOT [1], predictions were made for seven systems in CAPRI Rounds 1 and 2. DOT's performance is an important part of DoME's development, since DoME's energy model is meant to be a higher precision, continuous version of that

used by DOT. Ours was one of four teams (out of nineteen) submitting good predictions for three of the seven systems (the most achieved by any group.) This work will appear in a special edition of Proteins [2].

As a useful starting point for the development of DoME, we are running extensive analysis on the CAPRI systems. Part of the goal in developing DoME is to achieve accurate docking solutions without requiring as extensive a search as DOT performs. We hope to determine how fineness of the sampling affects the quality of the results, in particular at what sampling level DoME returns results with a high fraction of residue-residue contacts. The effect of local optimization of the best DoME solutions generated in the global scan is being considered, as well as local optimization (using DoME) of solutions generated by DOT's more exhaustive search. It appears that optimization can disrupt correct residue-residue contacts, but it is not yet clear whether the reason for this is biological or algorithmic. We are also implementing and testing schemes for global optimization. The aim is to find a consistent "recipe" of scanning, local and global optimization that will produce useful results for most protein-protein systems. Global optimization for docking presents some difficulties, as the variables used to parameterize the system are non-homogeneous and in some cases cyclic.

[1] J.G. Mandell, V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny and L.F. Ten Eyck (2001), "Protein docking using continuum electrostatics and geometric fit," *Protein Engineering* **14**(2): 103—115.

[2] D.H. Law, L.F. Ten Eyck, O. Katzenelson, I. Tsigelny, V.A. Roberts, M.E. Pique and J.C. Mitchell (2002), "Finding needles in haystacks: Re-ranking DOT results using shape complementarity, cluster analysis and biological information," *Prot. Struct. Fun. Gen.* In press.

A50

Functional Analysis and Discovery of Microbial Genes Transforming Metallic and Organic Pollutants: Database and Experimental Tools

Lawrence P. Wackett
(wackett@biosci.cbs.umn.edu) and Lynda B.M. Ellis (lynda@mail.ahc.umn.edu)

Center for Microbial and Plant Genomics, University of Minnesota

It is the major thesis of the current project that much of the breadth of microbial metabolism remains uncatalogued and uncharacterized. Characterizing this metabolism represents a major task of microbial functional genomics. Moreover, there is a general inability to predict metabolic pathways when all of the necessary reactions are not found in databases. The research described here seeks to better assemble existing metabolic data, discover new microbial metabolism, and predict metabolic pathways for compounds not yet in databases.

Approximately half the chemical elements are metallic and metalloid. Microbial metabolism of many of these elements, and compounds containing them, have been poorly studied relative to common intermediary metabolism, the typical focus of functional genomic analysis. Yet, recent studies suggest that many microbes have broad abilities to transform metals, metalloid elements, and compounds containing those elements. In the current project, the web-based University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) has been expanded to include information on microbiological interactions with 52 chemical elements. For each element, there is now a webpage with annotation on the major microbial interactions with that element, links to Medline, and access to further UM-BBD information. The information above has been coordinated with several depictions of the periodic table of the elements, one a classical table with columns and rows, and secondly a depiction of the elements in a spiral. The latter serves to cluster elements better with respect to their interactions with microbiological systems. The URLs for the relevant pages are given below:

- Biochemical Periodic Tables (overview website): http://umbbd.ahc.umn.edu/periodic/

- Traditional Periodic Table: http://umbbd.ahc.umn.edu/periodic/periodic.html
- Spiral Periodic Table: http://umbbd.ahc.umn.edu/periodic/spiral.html

The pages above have added hundreds of new linkages to UM-BBD compound pages. For example, the mercury element page has 4 links, arsenic has 9 links, and chlorine has 129 links to UM-BBD compounds, respectively.

An important facet of the current project is to discover new metabolism and functionally analyze the novel microbial enzymes and genes involved. A bioinformatic analysis has shown that on the order of one hundred chemical functional groups are found in natural products, yet only fifty are currently known to undergo microbial transformation. In this context, complete genome sequence annotation will require the identification of genes and enzymes that metabolize the full diversity of elements and functional groups that microbes act on in the environment. In the current project, we are uncovering the molecular basis of microbial metabolism, some of which has been previously unknown and uninvestigated. For example, we have investigated the metabolism of bismuth compounds, boronic acids, azetidine ring compounds, and novel organonitrogen compounds.

Another goal of the project has been to develop a tool to predict microbial catabolism, using the UM-BBD as a knowledge base. The objective is to propose one or more plausible biodegradation schemes for compounds whose metabolism is not yet known. To begin, a system for substructure searching was added to the UM-BBD, which both improved the database and was a necessary component for developing a metabolism prediction software. The metabolism prediction software is based on rules that describe fundamental microbial reactions. At present, there are 88 rules in the biotransformation rule database, each specifying the atoms and their positions in a functional group and the biotransformation reaction that they undergo. The software can now predict biodegradation pathways for a significant number of aliphatic and aromatic compounds. The prototype system has been offered to UM-BBD users to use and critique. The system will be expanded with input from our Scientific Advisory Board and the broader scientific community in the coming year.

A52

Comparative Genomics Approaches to Elucidate Transcription Regulatory Networks

Lee Ann McCue[1]* (mccue@wadsworth.org), William Thompson[1], C. Steven Carmack[1], Zhaohui S. Qin[2], Jun S. Liu[2], and **Charles E. Lawrence**[1]

*Presenter

[1]The Wadsworth Center, New York State Department of Health, Albany, NY 12201; and [2]Department of Statistics, Harvard University, Cambridge, MA 02138

The ultimate goal of this research is to delineate the core transcription regulatory network of a prokaryote. Toward that end, we are developing comparative genomics approaches that are designed to identify complete sets of transcription factor (TF) binding sites and infer regulons without evidence of co-expression. Using *Escherichia coli* as our model system, we have developed a phylogenetic footprinting technique to identify TF binding sites upstream of every operon in the *E. coli* genome. This method requires the genome sequences of several closely related species, and employs an extended Gibbs sampling algorithm to analyze orthologous promoter data. Using the promoters of 166 *E. coli* operons and a database of experimentally verified TF binding sites for validation, we have evaluated our ability to predict regulatory sites with this method, and addressed the questions of which species are most useful and how many genomes are sufficient for comparison. Orthologous promoter data from just three species were sufficient for ~75% of predicted sites to overlap the experimentally verified sites by 10 bp or more. A genome-scale phylogenetic footprinting study of *E. coli* identified 741 predictions above a threshold for statistical significance ($p < 0.05$) determined using randomized data simulations. We have also developed a novel Bayesian clustering algorithm to cluster these predictions thereby identifying 181 putative regulons, most of which are orphans—that is, the cognate TF is not known. This strategy to infer regulons utilizes only genome sequence information and is complimentary to and confirmative of gene expression data generated by microarray experiments. We are now applying these technologies to the *Synechocystis* PCC6803 genome.

A54

Predicting Genes from Prokaryotic Genomes: Are "Atypical" Genes Derived from Lateral Gene Transfer?

John Besemer[1], Yuan Tian[2], John Logsdon[1], and **Mark Borodovsky**[2] (mark@amber.biology.gatech.edu)

[1]Department of Biology, Emory University, Atlanta, GA; and [2]School of Biology, Georgia Technical Institute, Atlanta, GA

Algorithmic methods for gene prediction have been developed and successfully applied to many different prokaryotic genome sequences. As the set of genes in a particular genome is not homogeneous with respect to DNA sequence composition features, the GeneMark.hmm program utilizes two Markov models representing distinct classes of protein coding genes denoted "typical" and "atypical." Atypical genes are those whose DNA features deviate significantly from those classified as typical and they represent approximately 10% of any given genome. In addition to the inherent interest of more accurately predicting genes, the atypical status of these genes may also reflect their separate evolutionary ancestry from other genes in that genome. We hypothesize that atypical genes are largely comprised of those genes that have been relatively recently acquired through lateral gene transfer (LGT). If so, what fraction of atypical genes are such *bona fide* LGTs? We have made atypical gene predictions for all fully completed prokaryotic genomes; we have been able to compare these results to other "surrogate" methods of LGT prediction. In order to validate the use of atypical genes for LGT detection, we are building a bioinformatic analysis pipeline to rigorously test each of the gene candidates within an explicit phylogenetic framework. This process starts with gene predictions and ends with a phylogenetic reconstruction of each candidate. From the set of *bona fide* LGTs that we have identified, we will be able to determine the LGT parameters to which our gene finding programs are most sensitive (*i.e.* time scale of transfers, phylogenetic distance from transfer source, *etc*.). We are developing this pipeline using four cyanobacterial genomes as our test set: *Prochlorococcus marinus* str. MIT 9313, *Prochlorococcus marinus* subsp. Pastoris str. CCMP1378, *Synechococcus sp*. WH 8102, *Synechocystis sp*. PCC 6803, the first three of which are nearly complete genomes from DOE. From this initial analysis, we are estimating the extent and pattern of LGT in each of these genomes. We will then extend our studies to include all available genomes, both complete and nearly complete.

A56

Advanced Molecular Simulations of *E. coli* Polymerase III

Michael Colvin[1] (colvin2@llnl.gov), Felice Lightstone[1], Ed Lau[1], Ceslovas Venclovas[1], Daniel Barsky[1], Michael Thelen[1], Giulia Galli[2], Eric Schwegler[2], and Francois Gygi[3]

[1]Biology and Biotechnology Research Program, [2]Physics and Advanced Technology Directorate, and [3]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94552

The goal of this project is to use advanced molecular simulation methods to improve our understanding of several key mechanisms in the *E. coli* DNA polymerase III (Pol III). Pol III is the primary replicating polymerase in *E. coli* and the holoenzyme consists of at least 10 protein subunits that carry out different functions, including template-driven DNA replication (subunit α), replicative error correction (subunit e), tethering of the Pol III complex on the DNA by a "clamp" (subunit β), and the clamp loading complex (γ, δ, δ', ψ and χ subunits). In this poster we will present several key results from this study involving simulation methods ranging from homology-based structure prediction to first principles molecular dynamics simulations.

Although the three-dimensional structure of *E. coli* polymerase III β-clamp is known, and the structure of the clamp-loading complex has been solved recently, the mechanism of loading the β-clamp onto primed DNA sites remains unclear. One of the unanswered questions is what forces act to direct DNA into the β-clamp, once it becomes opened by the clamp-loader. The crystal structures of the β-clamp, clamp-loader and the monomer of the β-clamp complexed with one of the subunits (δ) of the clamp-loader makes it clear that there are no flexible domains/subdomains of any kind that could push DNA inside the ring-shaped β-clamp. On the other hand, the experimental evidence suggests that the ATP-dependent clamp-loading reaction is very efficient, arguing against a random diffusion-based mechanism.

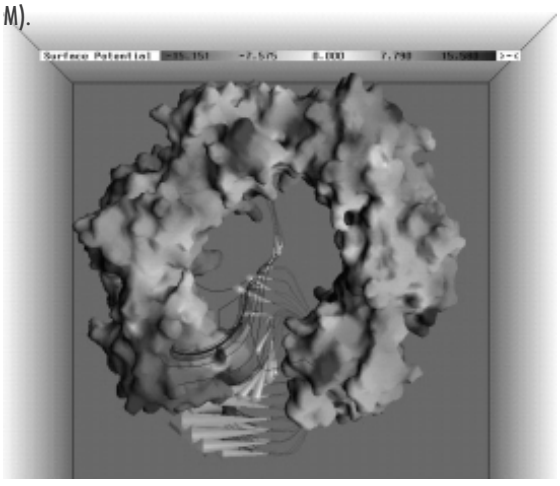
We hypothesized that the lack of the structural "helper" motifs might be compensated by other properties of the β-clamp itself, such as electrostatics. The crystal structure of a mutant β-clamp provided us with a model of the clamp opened at a single interface. Although the opening at the interface is too small for DNA to pass into the ring, we considered it a feasible model for the estimation of the effect of the electromagnetic field on the negative charges (DNA) in the vicinity of the opening. Using combination of DelPhi (a program to solve the non-linear Poisson-Boltzmann equation) and GRASP, we analyzed the electrostatic properties of the open β-clamp. The computational study revealed that the vectors of the electromagnetic field near the opened β-clamp interface are directed such that the negative electric charge (like DNA) would be drawn into the opening (see Figure). Interestingly, if the interface is opened very widely, the electrostatic guiding effect largely disappears. We are also performing classical molecular dynamics simulations of the β-subunit and its interactions with DNA.

The epsilon (ε) subunit is the 3'–5' exonuclease in this DNA polymerase and interacts with the α (polymerase unit) and θ (unknown function) subunits. Epsilon is able to hydrolyze the phosphodiester backbone of either single or double stranded DNA. This enzyme requires Mn^{2+} or Mg^{2+} for activity. The exonuclease activity resides in the N-terminus (residues 1-186). The C-terminus (residues 187-242) binds to the α subunit. The crystal structure of ε (residues 1-186) complexed with the p-nitrophenyl ester of TMP at pH 5.8 and 8.5 has recently been solved. These two structures have been provided by our external collaborator to serve as the starting point for this study. In this classical molecular dynamics study the interactions formed between a trinucleotide, modeled into the active site, and the ε subunit were investigated. Four molecular dynamics simulations were performed using explicit solvent molecules. In three separate simulations, the likely general base amino acid (His162) was modeled as a neutral residue (protonated at ND1 or NE2) and an ionized residue. The fourth simulation contained the quantum chemical gas-phase transition state docked into the active site and His162 was modeled as an ionized residue.

Our results show that the phosphodiester backbone interacts with the two Mg^{2+} ions and ε, the nucleobases form surprisingly few interactions with ε. G1 of the trinucleotide is almost completely solvent exposed. Only Met18, Asn99, and Phe102 consistently interact with the bases in all simulations. Glu61 and Ala62 interacted with the bases in the majority of the simulations. In contrast, the phosphate undergoing hydrolysis is highly stabilized in the active site. There is a minimal amount of motion in this group relative to the rest of the nucleotide. Only in the TS simulation does His162 stay close to the phosphate throughout the simulation. His162 is situated in a mobile loop which exhibits some of the highest fluctuations in the crystal structure.

In our efforts to understand the basic chemistry of the epsilon subunit (exonuclease), we have used first principles molecular dynamics simulations (FPMD) to simulate a model nuclease substrate, dimethyl phosphate (DMP) and the hydroxide-induced hydrolysis of DMP. These simulations were run on a massively parallel computer and the 14 ps DMP simulation is the longest FPMD ever reported. The results of this simulation show that the proximity of the solvated magnesium ion induces conformational changes in DMP. These subsequent conformations are higher energy conformations and could be a factor as how the enzyme cleaves the DNA backbone so easily. Hydroxide attack on DMP was simulated by fixing the distances over the reaction coordinate and then sampling for 3 ps for each constrained distance. The results show a small shoulder in the

Colvin—Fig. 1. Electrostatic surface of the slightly opened β-clamp. Lines correspond to the lines of the electric field. Arrows indicate directionality of the electric field, and their size is proportional to the calculated force. Calculations were made in solution, considering dielectric constant 2.0 for the protein interior and 80.0 for the solvent and the physiological ionic strength (0.15 M).



reaction free energy profile after the initial transition state as hydroxide attacks DMP. These results provide a reference system for subsequent simulations for the exonuclease-catalyzed reaction, and further are providing clues to answering the 40-year debate whether phosphate hydrolysis has a single transition state or has a stable intermediate.

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

A58

Karyote®: Automated Physico-Chemical Cell Model Development Through Information Theory

Peter J. Ortoleva (ortoleva@indiana.edu), Abdalla Sayyed-Ahmad, Ali Navid, Kagan Tuncay, and Elizabeth Weitzke

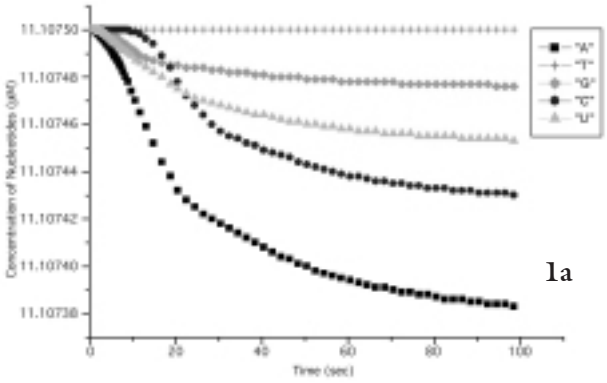
Center for Cell and Virus Theory, Indiana University

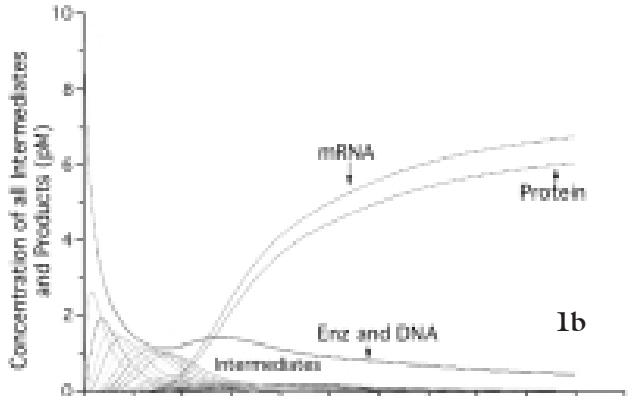
The dynamics of a cell are modeled using a reaction, transport, and genetic simulator, *Karyote*®, in order to predict cell behavior in response to changes in its surroundings or to modifications of its genetic code. Our methodology accounts for the organelles of eukaryotes and the specialized zones in prokaryotes by dividing the volume of the cell into discrete compartments. Each compartment exchanges mass with other compartments either through membrane transport or with a time delay effect associated with molecular migration (e.g. as for the nucleoid in prokaryotes). In each compartment multiple metabolic, proteomic and genomic reactions take place. All couplings among processes are accounted for. A multiple scale technique allows for the computation of processes that occur on a wide range of time scales.

Karyote® allows for the investigation of various cell behaviors that arise due to gene mutation, presence of external chemical agents or other factors. The underlying equations integrate the metabolic, proteomic and genetic networks (see Fig. 1). Catalyzed polymerization kinetics transcribes mRNA from an input DNA sequence while the resulting mRNA is used via ribosome-mediated polymerization kinetics to accomplish translation. Feedback associated with the creation of species necessary for metabolism by the genomic/proteomic network modifies the rates of production of factors (e.g. nucleotides and amino acids) that affect the genome/proteome dynamic. Hence, the effect of genetic mutations on overall cell behavior is accounted for. The concentration and sequence of the predicted proteins can be compared with experimental data via the construction of synthetic tryptic digests and associated mass spectra.

The complex network of biochemical reaction/transport processes and their biochemical spatial organization make the development of a predictive model of a living cell a grand challenge for the 21st century. However, advances in reaction/transport modeling and the exponentially growing databases of genomic, proteomic, meta-

Ortoleva—Fig. 1a and b. 1 (a) *Karyote* predicted time dependence of nucleotide concentrations during transcription for the gene TACTTTTAGGGG. As nucleotides are depleted, mRNA synthesis slows down, illustrating an important feature of coupled genome, proteome, metabolome dynamics captured in *Karyote*. (b) *Karyote* predicted evolution of concentration over time of DNA, RNA Polymerase II, all of their complexes involved in transcription/translation and mRNA created under the same condition as seen in Fig 1(a) (Weitzke and Ortoleva 2003).





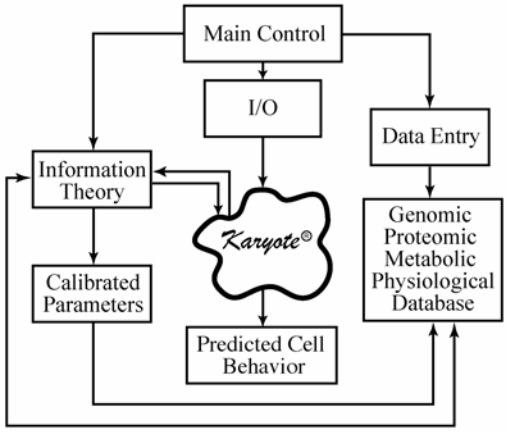


Fig. 2 This structure can serve as the basis for a national database wherein cell models are automatically developed/ calibrated and simultaneously used to interpret new data.

bolic and bioelectric data make cell modeling feasible if these two elements can be automatically integrated in an unbiased fashion. We have developed a procedure to integrate data with *Karyote®* using information theory (see Fig. 2).

Our procedure provides an objective approach for integrating a variety of types and qualities of experimental data. Data that can be used in this approach include NMR, spectroscopy, microscopy and cellular bioelectric information. The approach is demonstrated on the well-studied *Trypanosoma brucei* system.

A major obstacle for the development of a predictive cell model is that the complexity of these systems makes it unlikely that any model presently available will soon account for a complete set of processes. Thus, not only is the model-building endeavor labor intensive, but also at any stage one is faced with the challenge of calibrating and running an incomplete model. We present a probability functional method that allows the integration of quantitative and qualitative experimental data and physically motivated regularization to delineate the time course of the concentration of components (e.g. an enzyme) whose role may be key to the dynamics of the processes already incorporated in the model, but the reaction creating or destroying it are not yet understood.

A60

The Commercial Viability of EXCAVATOR™: A Software Tool For Gene Expression Data Clustering

Robin D. Zimmer[1]* (robzimmer@apocom.com), Morey Parang[2]*, Dong Xu[3], and Ying Xu[3]

*Presenters

[1]ApoCom Genomics, 11020 Solway School Road, Knoxville, TN 37931; and [2]Oak Ridge National Laboratory

ApoCom Genomics, in collaboration with Oak Ridge National Laboratory, is being funded under a DOE Phase I SBIR Grant (DE-FG02-02ER83365) to assess the commercial viability of a novel data clustering tool developed by Drs. Ying Xu, Victor Olman and Dong Xu (Xu, et.al., 2001). As we enter into an era of advanced expression studies and concomitant voluminous databases, there is a growing need to rapidly analyze and cluster data into common expression and functionality groupings. To date, the most prevalent approaches for gene and/or protein clustering have been hierarchical clustering (Eisen et.al., 1998), K–means clustering (Herwig et al., 1999), and clustering through Self-Organizing Maps (SOMs) (Tamayo et al.,1999). While these approaches have all clearly demonstrated their usefulness, they all have inherent weaknesses. First, none of these algorithms can, in general, rigorously guarantee to produce globally optimal clustering for any non-trivial objective function. Moreover K-means and SOMs heavily depend upon the 'regularity' of the geometric shape of cluster boundaries, and they generally do not work well when the clusters cannot be contained in some non-overlapping convex sets.

For cases where boundaries between clusters may not be clear, an objective function addressing more global properties of a cluster is needed. Three clustering algorithms, along with a minimum spanning tree (MST) representation, have been implemented within a computer program called EXpression data Clustering Analysis and VisualizATiOn Resource (EXCAVATOR™). Our research team has conducted a comparison between the EXCAVATOR™ clustering algorithm and the widely used K-means clustering algorithm using rat central nervous system (CNS) data. Two criteria were employed for the comparison. The first was based on the jackknife approach to assess the predictive power of the clustering algorithm,

and the second was based on the separability quality of clusters. All three of the EXCAVATOR™ algorithms (MST-hierarchical, MST-iterative, and MST-global optimal) outperformed the K-means algorithm relative to predictive power and separability quality.

In addition to comparative studies to assess the usefulness of EXCAVATOR™, the team has developed an advanced graphical user interface (GUI). The GUI has been designed to afford maximum flexibility incorporating the multi-clustering data visualization, as well as user driven comparison and editing capabilities. EXCAVATOR's™ data visualization component is based on a modular/flexible approach so as to extend its capability to other clustering/classification areas, such as phylogeny, sequence motif recognition, and protein family recognition.

References

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) *Cluster analysis and display of genome-wide expression patterns*. Proc. Natl Acad. Sci. USA, 95, 14 863-14 868.

Herwig, R., Poustka, A.J., Müller, C., Bull, C., Lehrach, H. and O'Brien, J. (1999) *Large-scale clustering of cDNA-fingerprinting data*. Genome Res., 9, 1093-1105.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation*. Proc. Natl Acad. Sci. USA, 96, 2907-2912.

Xu, Y., Olman, V. and Xu, D. (2001) *Clustering Gene Expression Data Using A Graph-Theoretic Appraoch: An Application of Minimum Spanning*. Bioinformatics. Vol.18, no. 2002.

A62

Modeling Electron Transfer in Flavocytochrome c_3 Fumarate Reductase

Dayle M. Smith, Michel Dupuis, Erich R. Vorpagel, and **T. P. Straatsma** (tps@pnl.gov)

Computational BioSciences Group, Biological Sciences Division, Pacific Northwest National Laboratory, P.O. Box 999, MS K1-83, Richland, WA 99352, (509) 375-2802, Fax (509) 375-6631

Ferric and ferrous hemes, such as those present in electron transfer proteins, often have low-lying spin states that are very close in energy. In order to explore the relationship between spin state, geometry and cytochrome electron transfer in the flavocytochrome c_3 fumarate reductase of *Shewanella frigidimarina*, we investigate, using density functional theory, the relative energies, electronic structure, and optimized geometries for a high- and low-spin ferric and ferrous heme model complex. Our model consists of an iron-porphyrin axially ligated by two imidazoles, which model the interaction of a heme with histidine residues. Using the B3LYP hybrid functional, we found that, in the ferric model heme complex, the doublet is lower in energy than the sextet by 8.4 kcal/mol, and the singlet ferrous heme is 6.7 kcal/mol more stable than the quintet. The difference between the high-spin ferric and ferrous model heme energies yields an adiabatic electron affinity (AEA) of 5.24 eV, and the low-spin AEA is 5.17 eV. Both values are large enough to ensure electron trapping, and electronic structure analysis indicates that the iron $d\pi$ orbital is involved in the electron transfer between hemes. Mössbauer parameters calculated to verify the B3LYP electronic structure correlate very well with experimental values. Isotropic hyperfine coupling constants for the ligand nitrogen atoms were also evaluated. The optimized geometries of the ferric and ferrous hemes are consistent with structures from X-ray crystallography and reveal that the iron-imidazole distances are significantly longer in the high-spin hemes, which suggests that the protein environment, modeled here by the imidazoles, plays an important role in regulating the spin state. Iron-imidazole dissociation energies, force constants and harmonic frequencies were calculated for the ferric and ferrous low-spin and high-spin hemes. In both the ferric and ferrous cases, a single imidazole ligand is more easily dissociated from the high-spin hemes.